September 9, 1997

RESEARCH AREAS

Prepared by
Michael D. McKay
Statistics Group, Los Alamos National Laboratory
Los Alamos, NM 87545

The research topics discussed below pertain to input uncertainty, model (structural) uncertainty, and to distinguishing between the two when comparing model prediction with external data. The topics address limitations and deficiencies of currently available analysis methods. The product of the research would be suitable methodologies or approaches in light of identified limitations and deficiencies, or reasons why no satisfactory solutions were found.

Input Importance

Variance-based analysis methods provide effective importance indicators. The indicators are used to evaluate individual inputs as well as subsets of inputs. Dominant input subsets, those which account for a substantial part of prediction uncertainty, can be constructed sequentially beginning with dominant single inputs. When the number of computer runs is limited, the process can be inefficient or imprecise by including more inputs than necessary in dominant subsets. Even when there is no limit on the number of computer runs, the time required to identify minimal subsets may be prohibitive. A research program investigating more precise and efficient methods in this area could be composed of the following five topics.

1. **Practical variance decompositions.** The ability to identify input subsets with conditional and partial variances has been demonstrated (McKay, 1995). The variance decompositions, however, are intrinsically of a binary nature in that they have two terms corresponding to a partition of model inputs into two sets. An example of a more general decomposition is one which has a term for each model input. Decompositions such as these usually, though not necessarily, depend on linearity assumptions and are, therefore, of limited value. It is proposed that conditional and partial variances and others be developed into more general, nonparametric variance decompositions. Research might follow along the lines of decompositions of Baybutt and Kurth (1978) and Cox (1982) or be developed from the decompositions suggested by Stein (1987) and Owen (1994).
2. **Experimental designs for estimation of variance components.** Variance decompositions require efficient experimental designs for estimation of the components of variance for the methods to be feasible. It is proposed that traditional variance component estimation techniques along with methods related to computer experiments be investigated. In particular, orthogonal arrays and extensions to LHS following Owen (1992) would be investigated.
3. **Smart variable selection procedures.** When essentially unlimited computer runs are feasible, brute-force sequential selection procedures may be adequate to select dominant input subsets. However, the number of computer runs necessary increases geometrically as the number of inputs increases, so that it quickly becomes infeasible to make the computer runs necessary for adequate model analysis. It is proposed that smart variable selection procedures

which take advantage of particulars of the variance decomposition and experimental design be developed. Investigation would begin with optimal procedures for subset selection in regression following Hocking (1967) and others.

4. **Simultaneous treatment of outputs.** The analysis of several outputs simultaneously with sequential methods can result in importance masking of some inputs for some outputs. Because many models have several outputs and require much computer time, successful research might make possible otherwise unattempted analyses. A starting point for research is the extensive literature of multivariate analysis.

5. **Analysis of simulation (stochastic) models.** A simulation model is one in which the input vector x contains parameters of the "stochastic" probability distribution of y. In MACCS, for example, the stochastic distribution of the output arises from randomly selected weather conditions. In past MACCS studies, the stochastic variability has been assumed to be treated adequately by analyzing CCDFs, with the CCDFs based on 100 weather samples. The validity of this approach, particularly when fewer than 100 samples are available, needs to be investigated. Investigations of this topic would begin with development of a variance-based mathematical formulation of the problem from which further research would proceed. In particular, methods to efficiently allocate computer runs between quantifying stochastic variability and identifying important inputs would be developed.

Characterization of the Space of Models

A sampling approach is usually taken when evaluating prediction uncertainty arising from input uncertainty. The space of plausible input values and a probability distribution defined on the space are used to obtain a sample of input values. The variability in model predictions for the sample of values is used to estimate the prediction uncertainty. By way of a parallel approach to model uncertainty, a space of plausible models and a probability distribution defined on the space are required. The definition of such a space and probability distribution are conceptually much more difficult for models than for input values. Research investigation related to the characterization of the space of models might begin with the work of Sacks, Welch, Mitchell and Wynn (1989) who represent models as realizations of stochastic processes. Simplifying (and possibly unrealistic) assumptions for stochastic processes underlie the fields of time series analysis and geostatistics, for example. How such assumptions might apply to viable model spaces for NRC applications needs to be examined. Research in this area is very much a wide open endeavor with an unknown probability of successful completion. Because of the importance NRC places on model uncertainly and because of the dearth of methods, payoffs for successes are very large: they would open the door to development of methods for assessing model uncertainty.

Distinguishing Components of Error

From a practical point of view, distinguishing between input value specification and model structure as the cause of incorrect model prediction can be very important to proper diagnosis of model prediction error. A possible approach was discussed in McKay (1995). Research could begin with that discussion. It would be expanded from there to include vector valued model outputs and multiple outputs.

References

Baybutt, P. and Kurth, R. E. (1978). Uncertainty analysis of light-water reactor meltdown accident consequences: Methodology development. Technical report, Report from Battelle's Columbus Laboratories to the U.S. Nuclear Regulatory Commission, Available from the author.

Cox, D. C. (1982). An analytical method for uncertainty analysis of nonlinear output functions, with application to fault-tree analysis. *IEEE Transactions on Reliability*, R-31(5):465–468.

Hocking, R. R. and Leslie, R. N. (1967). Selection of best subsets in regression analysis. *Technometrics*, page 531.

McKay, M. D. (1995). Evaluating prediction uncertainty. Technical Report NUREG/CR–6311, U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory.

Owen, A. B. (1992). Orthogonal arrays for computer integration and visualization. *Statistica Sinica*, 2(2):439–452.

Owen, A. B. (1994). Controlling correlations in latin hypercube samples. *Journal of the American Statistical Association*, 89(428):1517–1522.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435.

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2):143–151.